תרגיל בית – דאטה סיינס/ genAI

חברת ONE

1. רקע ומטרה

בתהליך ניתוח מסמכי מכרז, אחת הבעיות המרכזיות היא זיהוי מהיר ומדויק של מידע רלוונטי מתוך עשרות/מאות עמודים. תהליך ידני כזה לוקח בין שעות לבין שבועות כתלות במורכבויות שונות (אורך המכרז, פרמטרים מחולצים ועוד..)

כדי לעזור להם לענות על זה אוטומטית רצינו לפתור זאת על ידי מודלי שפה. במקום לשלוח את כל המסמך למודל שפה (בפעם אחת)– פעולה שמובילה לעלויות גבוהות, ירידה בדיוק ובזבוז משאבים – אנחנו מפרקים את המסמך לעמודים (או מקטעים), מתייגים את התוכן של כל עמוד מראש, ושולחים רק את החלקים הרלוונטיים עבור כל פרמטר.

מדובר בגישה מקבילה לתהליך(RAG (Retrieval-Augmented Generation , אך במקום לבצע חיפוש וקטורי על דאטה חתוך מראש, כאן התהליך כולל חיתוך, תיוג והתאמה בשלב הריצה (ולא האימון/הכנת הדאטה בייס)– על בסיס מטא-דאטה, כללים וניתוח מבני של המסמך.

מטרת התרגיל היא:

- לעבור על המסמך ולתייג אותו לפי עמודים/פרמטרים רצויים לחילוץ
- לזהות נושאים ותכנים לפי רשימת פרמטרים נתונה
- ולהתאים בין כל פרמטר לעמודים הרלוונטיים בלבד – כך שהמודל יעבוד רק על התוכן הנחוץ
- ואז לשלוח באמצעות פרומפט חכם את העמודים הרלוונטיים לחילוץ הערך המבוקש

2. קלט התרגיל

לצורך ביצוע התרגיל יסופקו שני קבצים:

1. קובץ PDF של מכרז אמיתי לדוגמה – לדוגמה tender_sample.pdf :
המסמך כולל מידע הרבה מידע על דברים שונים כולל על המזמין, תיאור השירות, תנאי סף, מועדים, קריטריונים להערכה, ערבות ועוד.
2. קובץ JSON ובו רשימת פרמטרים שעליך לאתר במסמך (הפרמטרים קבועים למשימה זו).
המטרה היא לזהות אילו עמודים במסמך רלוונטיים עבור כל פרמטר.
אם פרמטר לא מופיע במסמך – יש לציין במפורש "לא נמצא" (string). יש פרמטר/ים שאינם קיימים במסמך שנכללו במכוון, כדי לבדוק טיפול במקרים כאלה.

3. משימות

עליך לפתח רכיב תוכנה שמקבל כקלט את קובץ המכרז ואת רשימת הפרמטרים, ומחזיר עבור כל פרמטר את רשימת העמודים במסמך שבהם סביר להניח שנמצא המידע הרלוונטי או שכדאי לשלוח אליהם את הפרמטר ובאמצעות זה את ערך הפרמטר עם תוספות לפלט שנפרט בהמשך.

אנחנו מאוד מעריכים חשיבה מודולרית ופירוק המשימה לפונקציות ולא רק סקריפט אחד ארוך!!

<u>שלבים:</u>

חיתוך המסמך לעמודים

קריאה וחיתוך לעמודים של המסמך - כל עמוד יטופל כיחידה נפרדת לניתוח.

תיוג כל עמוד לפי תוכנו

- יש לזהות עבור כל עמוד אילו פרמטרים (אם בכלל) מופיעים בו בפועל.
- בנוסף, ניתן להוסיף תיוג אחד (מהבאים) נוסף לפי שיקולך:
  - נושא משוער של העמוד לפי מילות מפתח
  - תיוג אחר שיכול לסייע בהתאמת עמודים לפרמטרים
- מומלץ להשתמש במודל שפה חכם כדי לעלות דיוק בזמן פיתוח קצר שיש למשימה - אך לא חייב.
- יש לוודא איתור איכותי גם במקרים של: מילים נרדפות, הטיות לשון, ניסוחים עקיפים, תרגום או שימוש מונחים שונים למסר דומה
- המלצה להשתמש פה בפרומפט שתופס אותם!

התאמה בין פרמטרים לעמודים

לאחר מכן, עבור <u>כל פרמטר</u>, יש למפות את העמודים הרלוונטיים במסמך. למשל:

my_parameter_pages =[2,3,7,8,40]

אם פרמטר לא נמצא כלל – יש לציין זאת במפורש בפלט כ "לא נמצא" (string).

הפלט צריך להתאים לפורמט הצנרת הקיימת (בהמשך תופיע דוגמה).

בניית פרומפט בסיסי לכל פרמטר

- כתוב פרומפט עבור כל פרמטר, שמטרתו להנחות את מודל השפה לחלץ את המידע מהעמודים שנבחרו.
- רצוי לבנות את הפרומפטים בצורה מודולרית שתאפשר שימוש חוזר על פני מספר פרמטרים דומים (חשוב!!).

שליחת הפרומפט למודל ושמירת התגובה

- הפעל את הפרומפטים מול מודל שפה שוב – הפעם כדי לקבל את הערך המחולץ (עדיף דרך api למודל חכם, אך לא חובה).
- בעצם לכל פרמטר לשלוח את הפרומפט ואת העמודים הרלוונטיים לו
- לשם פשטות וחיסכון בזמן – אפשר להתעלם מטיפול בטעויות ובעיות בחזרה ב api
- שמור את הפלטים במבנה קבוע:
  - ראו דוגמא בהמשך
  - עדיף בתוך קובץ או שאם הקידוד בעברית לוקח זמן (מבחינת לשמור בוורד או פידיאף) אפשר להדפיס משתנה בקונסול/דיבאגר/notebook

רשימת הפרמטרים עם הסברים

להלן 7 פרמטרים שנבחרו לתרגיל. עליכם לזהות עבור כל אחד מהם את העמודים הרלוונטיים במסמך.
בבקשה לא להעתיק מפה אלא לקרוא את הקובץ ג׳ייסון ולקחת את הרשימה משם.

הסבר	שם בפרויקט (אנגלית)	שם הפרמטר (עברית)
שם הגוף שפרסם את המכרז (למשל: רשות, תאגיד, עירייה)	client_name	שם המזמין
שם מלא של המכרז, כולל מספר וסוג (פומבי, דו-שלבי וכו)	tender_name	שם המכרז
דרישות חובה להשתתפות – לדוגמה: ניסיון, עמידה בחוקים, רישוי.	threshold_conditions	תנאי סף
כמה זמן תימשך ההתקשרות, והאם קיימות אופציות להארכה.	contract_period	תקופת ההתקשרות
כיצד נשקלל את ההצעות – מחיר מול איכות, קריטריונים לניקוד.	evaluation_method	שיטת ההערכה (מפ״ל)
סכום הערבות, סוג הערבות, תנאים לפירעון, תוקף.	bid_guarantee	ערבות מכרז
פרמטר לא רלוונטי שנוסף בכוונה – אין לצפות שימצא במסמך. יש לציין "לא נמצא" (string). לא לרשום זאת למודל, אלא אמורים לשלוח פורמפט ודפים כרגיל	idea_author	הוגה הרעיון

5. מבנה הפלט המצופה

עליך להחזיר עבור כל פרמטר את המידע הבא:

- שם הפרמטר
- ערך הפרמטר -הערך המרכזי שחולץ מהמסמך
- פרטים - פירוט נוסף, ניסוח מורחב או פרשנות
- מקור- העמוד או המיקום ממנו נלקח המידע (יכול להיות רצף עמודים או מספר עמודים שונים או עמוד יחיד)
- ציון - ציון כולל (1–5) לאיכות החילוץ– לבקש מהמודל שיעשה זאת -

דוגמה לפורמט הפלט:(JSON)


```
{
  "client_name": {
    "answer": "רשות המים והביוב",
    "details": "תאגיד עירוני שהוקם לפי חוק תאגידי מים וביוב",
    "source": "עמוד 2, פסקה ראשונה",
    "score": 5
  },
  "tender_name": {
    "answer": "[redacted] – 28/2024 'מכרז מס",
    "details": "[redacted]",
    "source": "עמוד 1, כותרת",
    "score": 4
  },
  "[redacted]": {
    "answer": "",
    "details": "",
    "source": "לא נמצא",
    "score": 0
  }
}
```


<u>הנחיות:</u>

- הפלט חייב להתבסס על **תוכן אמיתי מתוך המסמך**.
- אם פרמטר לא מופיע:
  - – השאר את answer ו־ details ריקים
  - Source- לא נמצא
  - ציון – לפי רמת הוודאות של המודל שהוא בטוח שהוא לא מצא

- בספריית זיפ ששלחנו (מבחן בית – one) יש כמה קבצים:
  - המסמך הזה

- פרמטרים לחילוץ
- מכרז ללימוד עם תשובות
- תשובות של מכרז (רק להשראה, אין צורך להשתמש בו)
- מכרז נוסף לבדיקת הקוד (סוג של test)

- הדגש הוא על מבנה הפיתרון, חשיבה עמוקה ולא בהכרח על הדיוק.
- מודולריות זה חשוב – בנו מספר פונקציות שקל להחליף להם פרמטרים.
- חשוב מאוד!!

למען הפשטות שלחנו מסמכים קצרים. במצב כזה אפשר גם בקלות לשלוח כל פעם את כל המסמך למודל שפה ולקבל תוצאות – זו לא כוונת המשורר – עשינו זאת כדי להקל – ראו בשליחת מינימום עמודים לכל פרמטר כאחת המטרות. במציאות מכרז גם יכול להיות 800 עמודים

דרישות טכניות:

- שפת תכנות: פייתון
- ספריות/מודולים מותרים: כל מודל
- מודלי שפה לשימוש: מה שהכי מהיר וטוב לכם
- דרישות מבנה הקוד:

- יש להגיש **סקריפט אחד או מחברת Jupyter אחת** שמריץ את כל התהליך מקצה לקצה
- לחילופין סרטון אמיתי **<u>ולא ערוך</u>** של הכל רץ אצליכם עם הסברים על הפונקציות השונות
- הקוד צריך לכלול:
  - קריאת המסמך
  - ניתוח לפי עמודים
  - תיוג
  - התאמת פרמטרים לעמודים
  - יצירת פלט לפי המבנה שנקבע
- חשוב מאוד!!!

תיהיה עדיפות (כמו שתואר כבר) לחשיבה מודלרית עם פונקציות. משמע איזו פונציית main שקוראת לפונציות קטנות/בינוניות.

למשל:

```
• def main():
      print("starting… ")
      init_time = time.time()

      load_data(file_name=file_name,….)
    o  more functions here
```